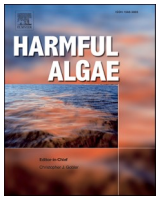
Harmful Algae 137 (2024) 102679

Contents lists available at ScienceDirect

Harmful Algae

journal homepage: www.elsevier.com/locate/hal



An evaluation of statistical models of microcystin detection in lakes applied forward under varying climate conditions

Grace M. Wilkinson [a,*], Jonathan A. Walter [b], Ellen A. Albright [a], Rachel F. King [c], Eric K. Moody [d], David A. Ortiz [a]

[a] Center for Limnology, University of Wisconsin – Madison, 680N Park Street, Madison, WI 53706, USA
[b] Center for Watershed Sciences, University of California – Davis, One Shields Ave., Davis, CA 95616, USA
[c] Department of Ecology, Evolution, and Organismal Biology, Iowa State University, 2200 Osborne Dr., Ames, IA 50011, USA
[d] Department of Biology, Middlebury College, Middlebury, VT 05753, USA

ARTICLE INFO

Edited by: Dr Assaf Sukenik

Keywords:
Microcystin
Cyanobacteria
Lake
Reservoir
Harmful algal blooms
Cyanotoxins

ABSTRACT

Algal blooms can threaten human health if cyanotoxins such as microcystin are produced by cyanobacteria. Regularly monitoring microcystin concentrations in recreational waters to inform management action is a tool for protecting public health; however, monitoring cyanotoxins is resource- and time-intensive. Statistical models that identify waterbodies likely to produce microcystin can help guide monitoring efforts, but variability in bloom severity and cyanotoxin production among lakes and years makes prediction challenging. We evaluated the skill of a statistical classification model developed from water quality surveys in one season with low temporal replication but broad spatial coverage to predict if microcystin is likely to be detected in a lake in subsequent years. We used summertime monitoring data from 128 lakes in Iowa (USA) sampled between 2017 and 2021 to build and evaluate a predictive model of microcystin detection as a function of lake physical and chemical attributes, watershed characteristics, zooplankton abundance, and weather. The model built from 2017 data identified pH, total nutrient concentrations, and ecogeographic variables as the best predictors of microcystin detection in this population of lakes. We then applied the 2017 classification model to data collected in subsequent years and found that model skill declined but remained effective at predicting microcystin detection (area under the curve, AUC $\geq$ 0.7). We assessed if classification skill could be improved by assimilating the previous years' monitoring data into the model, but model skill was only minimally enhanced. Overall, the classification model remained reliable under varying climatic conditions. Finally, we tested if early season observations could be combined with a trained model to provide early warning for late summer microcystin detection, but model skill was low in all years and below the AUC threshold for two years. The results of these modeling exercises support the application of correlative analyses built on single-season sampling data to monitoring decision-making, but similar investigations are needed in other regions to build further evidence for this approach in management application.

1. Introduction

Earth's changing climate, land use intensification, and widespread nutrient enrichment (Stoddard et al., 2016) are altering the frequency and intensity of algal blooms in some, but not all lakes (Ho et al., 2019; Wilkinson et al., 2022). Variability in bloom severity among lakes and across years makes year-to-year prediction difficult (Beal et al., 2023; Rousso et al., 2020). Similarly, predicting whether a cyanobacterial bloom will produce cyanotoxins at concentrations high enough to threaten human, pet, and livestock health remains challenging given the variability in cyanotoxin production among lakes and years (Beversdorf et al., 2015; Gorney et al., 2023). Human exposure to microcystin, one of the most prevalent cyanotoxins in inland waters (Rastogi et al., 2014), can result in headaches and gastrointestinal symptoms (Carmichael and Boyer, 2016), with higher or chronic exposure being linked to colorectal cancer, liver damage, and in some cases, death (de Figueiredo et al.,

Open Research Statement: Collated data sets and analysis scripts used for this study are available at https://doi.org/10.5281/zenodo.12005095.

* Corresponding author.
E-mail address: gwilkinson@wisc.edu (G.M. Wilkinson).

https://doi.org/10.1016/j.hal.2024.102679
Received 14 March 2024; Received in revised form 10 June 2024; Accepted 15 June 2024
Available online 17 June 2024

2004). Regular monitoring of microcystin concentrations in recreational waters to inform management action is a tool for protecting public health; however, monitoring cyanotoxins is both resource- and time-intensive, which can leave communities near waterbodies at risk if cyanotoxins are insufficiently monitored.

The likelihood of occurrence and magnitude of an algal bloom is shaped by the complex interaction of in-lake conditions, watershed characteristics, and climatic drivers (Rousso et al., 2020; Taranu et al., 2017). Cyanobacterial blooms that produce high concentrations of microcystin most frequently occur in eutrophic lakes at low nitrogen to phosphorus (N:P) ratios (Harris et al., 2014; Orihel et al., 2012). Nutrient-rich lakes with high phytoplankton biomass generally have watersheds dominated by agriculture or urban land uses (Arbuckle and Downing, 2001; Beaver et al., 2014). Availability of N in particular is tied to cyanobacteria biomass and microcystin concentrations in eutrophic lakes, likely because of microcystins being an N-rich group of molecules (Beversdorf et al., 2015; Gobler et al., 2016; Van De Waal et al., 2014; Wagner et al., 2021). In addition to nutrient availability, blooms are regulated by a combination of light availability, water temperatures, stratification, and zooplankton grazing (Carpenter et al., 2022a; Reinl et al., 2023; Rousso et al., 2020). Interannual variation in climatic drivers such as precipitation and temperature can influence in-lake conditions that either suppress or promote cyanobacteria dominance and cyanotoxin production. The magnitude, stoichiometry, and timing of nutrient loading to lakes varies with precipitation (Kincaid et al., 2020), with extremes in precipitation linked to extremes in cyanobacteria biomass (Carpenter et al., 2022b). Similarly, when comparing lakes at a continental scale, warmer spring temperatures have been linked to higher cyanobacteria biomass but lower microcystin concentrations (Ho and Michalak, 2020).

There have been several calls for ecology to become a more predictive science (e.g., Clark et al., 2001; Dietze et al., 2018), emphasizing the benefits of prediction for applications to conservation and ecosystem management, and as a strong test of conceptual understandings. Despite a growing literature on predicting and forecasting ecological phenomena (Aboal et al., 2005; Lofton et al., 2023; Rousso et al., 2020; Walter et al., 2023; Wheeler et al., 2024), the extent to which correlative models can yield useful predictions has not been widely evaluated. As one example, large-scale single sample 'snapshot surveys' of lakes have provided valuable insight into the in-lake, watershed, and climatic conditions correlated with microcystin detection at regional to continental scales (Beaver et al., 2014; Ho and Michalak, 2020; MacKeigan et al., 2023; Taranu et al., 2017). These surveys leverage broad environmental gradients to detect drivers of microcystin concentration among lakes. A frequently stated goal of large-scale microcystin correlation analyses is to allow managers to identify bloom drivers and lakes likely to have toxic blooms based on more easily monitored water quality variables; however, the predictive ability of these statistical models is rarely evaluated.

Given the potentially substantial time and financial resources needed to test for microcystin across the portfolio of recreational waters a small private or government agency oversee, correlative predictors can help direct limited resources for monitoring. However, large-scale snapshot surveys are not necessarily designed to capture the dynamic nature of microcystin production and cycling (Shingai and Wilkinson, 2023), which could potentially lead to the misclassification of lakes and misdirection of resources. Additionally, if the survey design does not encompass the full range of conditions for important environmental drivers, statistical relationships among these variables and microcystin may be weak or misleading and may therefore not be informative to managers. A snapshot survey conducted in a single year will reflect the unique climatic conditions of that year and may therefore generate particular statistical relationships between drivers such as nutrient loading (Carpenter et al., 2022b; Loecke et al., 2017) and microcystin, which may not be consistent in years with differing climatic conditions. For example, extreme precipitation events that follow long dry periods generate larger loads of N relative to when events of similar magnitude occur following wetter conditions (Loecke et al. 2017). The utility and limitations of applying statistical models developed from surveys with low temporal replication (e.g., one sampling event) but broad spatial coverage to predict if lakes are likely to produce microcystin have not been fully evaluated.

We used five years of summertime water quality monitoring data from 129 lakes in Iowa (USA) to build and test a machine learning classification model for microcystin detection. The goal of this modeling effort was to evaluate the accuracy and limitations of applying a microcystin detection model based on one year's data to subsequent years for a large population of lakes in a highly modified landscape. Iowa has the largest area of land in row crop agriculture of any state in the US, and as a result the lakes and reservoirs of the state are highly enriched in N and P (e.g., Table 1) (Arbuckle and Downing, 2001; Filstrup and Downing, 2017). Cyanobacteria biomass is high in many of Iowa's lakes and reservoirs (Filstrup et al., 2014) with microcystin frequently detected above state drinking water and recreational exposure thresholds (Fig. 1). We used routine monitoring data of physical, chemical, and biological variables collected during the summer of 2017 to construct a machine learning classification model for microcystin detection. From this model we were able to identify the best predictors of microcystin detection in lakes embedded in this nutrient-rich landscape. We then applied the 2017 classification model to monitoring data collected in subsequent years to evaluate model skill under varying climate conditions. We assessed if classification skill could be improved by assimilating the previous years' monitoring data into the trained model. Finally, we used early summer monitoring data in the trained classification model to predict late summer microcystin detection. This was done to test if early season monitoring observations could be combined with a trained model to provide early warning for late summer microcystin detection. Overall, these modeling exercises allowed us to evaluate the utility and accuracy of using correlative models built from snapshot survey monitoring data to provide support to decision makers regarding microcystin monitoring and public health protection.

2. Methods

2.1. Field sampling and laboratory analysis

The study lakes (Fig. 1b) are all publicly owned waterbodies in the state of Iowa, and are designated primarily for recreational use (although some supply drinking water), and are a part of the Iowa

Table 1
Select characteristics and measurements in the surface water (0–2 m) of the population of study lakes ($n = 67$–129 per year) located in Iowa, USA from 2017 to 2021.

Variable	Min	25th Percentile	50th Percentile	75th Percentile	Max
Surface Area (ha)	4	12.5	32	117	4452
Watershed Area (ha)	4	277	779	2723	142,190
% area of watershed as Cropland	0.0	23.8	45.8	60.4	89.2
Maximum Depth (m)	2.0	4.9	6.3	8.5	42.3
Total P (μg L^{-1})	7.8	50.9	82.7	134.8	975.7
Total N (mg L^{-1})	0.3	1.1	1.6	2.3	9.7
N:P ratio (molar)	5.5	29.0	41.8	61.8	898.4
Chlorophyll a (μg L^{-1})	1.0	9.2	19.7	44.3	301.4
pH	7.4	8.1	8.3	8.5	9.4
Total Dissolved Solids (mg L^{-1})	85.0	166.4	211.5	265.2	576.0
Zooplankton (mg L^{-1})	1.9	47.4	106.6	207.9	3323.8

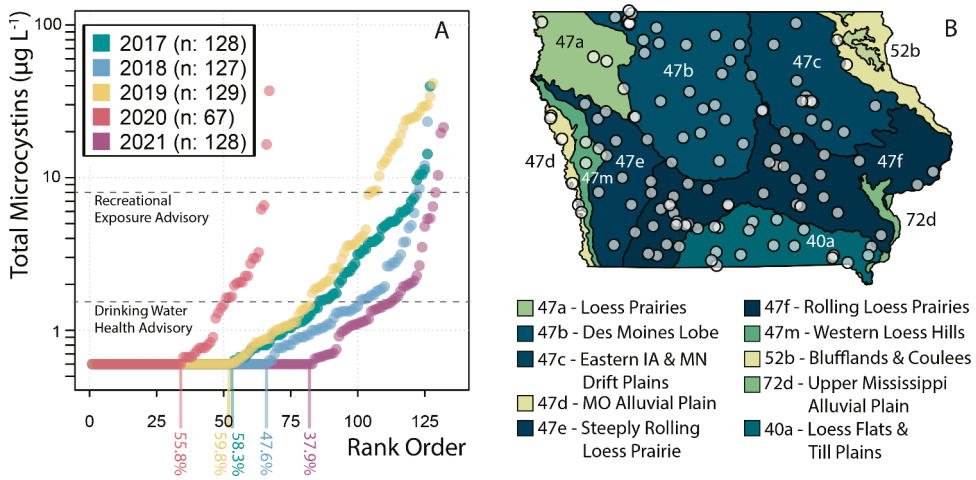


Fig. 1. (A) Mean total microcystins concentration by lake in each year. The concentrations are ranked in order from lowest to highest by year to show the distribution of values. The values along the horizontal axis are the percentage of lakes that year with detectable microcystin (>0.6 µg L^{-1}). Advisory thresholds for recreation and drinking water set by the US Environmental Protection Agency are noted (some, but not all lakes in the data set are drinking water sources). Fewer lakes were sampled in summer 2020 because of COVID-19 restrictions. (B) Map of the study lakes (semi-transparent white circles) and US Environmental Protection Agency level IV ecoregions.

Department of Natural Resources Ambient Lake Monitoring program. The waterbodies are a mix of natural lakes, impoundments, and filled borrow pits and quarries. Through the Ambient Lake Monitoring program, all lakes were sampled three times each between mid-May through the end of September from 2017 to 2021 with six weeks between each sampling event. Only 67 lakes were sampled in 2020 due to personnel and laboratory limitations from the COVID-19 pandemic.

At a location above the deepest point in each lake, a profile of water temperature, dissolved oxygen, pH, turbidity, and specific conductance was measured every 0.5 m with a YSI ProDSS handheld multiparameter sonde (YSI, Yellow Springs, Ohio). Secchi depth was also measured. The sensors were calibrated weekly except for dissolved oxygen which was calibrated daily. A column sampler was used to take an integrated water sample from the surface to the top of the thermocline when present (defined as $a > 1$° C decrease in temperature over <1 m and beginning at >1 m depth) or to 2 m depth, whichever was shallowest. The water from the integrated column sample was placed into a bucket rinsed in the matrix of the sample water, and then dispensed into sample bottles and placed in a cooler for transport. Water from the integrated column sample was analyzed for total microcystins and nodularins, phytoplankton biovolume, suspended solids, pigments (chlorophyll a and phycocyanin), total and inorganic nutrients, and alkalinity. Water samples for total microcystins were transferred to PETG amber vials with Teflon-lined caps to minimize the sorption of microcystin to the sample bottle. Phytoplankton samples were preserved with Lugol's solution and stored in amber bottles. Zooplankton were sampled by vertically towing a Wisconsin net with 63 µm mesh from the top of the thermocline to the surface, or from 0.5 m above the sediment surface if no thermocline was present. The zooplankton samples were preserved in the field with formalin and then transferred to 70 % ethanol after 24 h. All samples were kept on ice and in the dark until returning to the laboratory.

Sample preservation methods, holding times, quality control methods, and long-term method detection limits for each year are reported in Appendix 1, Table S1. Briefly, alkalinity was measured using end point titration. Total P and total Kjeldahl N were analyzed spectrophotometrically following digestion. Dissolved inorganic nutrient samples (orthophosphate, nitrate + nitrite, and ammonium + ammonia) were filtered through a Whatman glass fiber (GF) filter (pore size 0.45 µm) and preserved with sulfuric acid prior to analysis. Orthophosphate and ammonium + ammonia were measured spectrophotometrically as was nitrate + nitrite following cadmium reduction using a Seal AQ2 Discrete analyzer. Total N was calculated as the sum of total Kjeldahl N and nitrate + nitrite (Stanley et al., 2019). Chlorophyll-a and phycocyanin samples were filtered onto Whatman GF filters (pore size 1 µm), frozen, extracted in acetone and sodium phosphate buffer, respectively, sonicated, and measured fluorometrically. Suspended solids were measured by filtering water through a pre-weighed 934-AH GF filters (pore size 1.5 µm), drying and reweighing the filter, then combusting at 500 °C and reweighing the filter.

Total microcystins and nodularins (hereafter, total microcystins) were measured via the enzyme-linked immunosorbance assay method (ELISA) using kits from Gold Standard Diagnostics, following the USEPA 546 method. In 2017, the project reporting limit for total microcystins was set at 0.6 µg L^{-1}. Despite a decrease in laboratory reporting limits in later years, the initial project reporting limit of 0.6 µg L^{-1} was maintained for all years of data collection for the purpose of this statistical analysis. This reporting limit lies within the range or below the thresholds used by the state of Iowa for enhanced raw water monitoring (0.3–5 µg L^{-1}), finished drinking water advisories (1.6 µg L^{-1}), and recreational exposure advisory limit (8 µg L^{-1}).

Phytoplankton cells were concentrated through settling to a target of 30 natural units in each view field at 400x magnification. A subsample was placed in a Palmer-Maloney style nanoplankton chamber for identification and enumeration of all natural units in a minimum of eight view fields and 300 natural units. The dimensions of the first 50 natural units for each genus were measured in addition to the individual cells of colonies and filaments. Biovolume per liter was calculated based on phytoplankton shape and then converted to wet biomass per liter assuming a 1:1 ratio of wet mass and biovolume (Hillebrand et al., 1999; Holmes et al., 1969). Zooplankton and phytoplankton samples were identified and enumerated to the lowest taxonomic unit possible using light microscopy. For zooplankton, a 1 mL subsample of the vertical net tow was enumerated and identified using a dissecting scope to genus for Cladocera and Rotifera and to order for Copepoda. Zooplankton biomass was calculated for each taxonomic group using allometric equations (Dumont et al., 1975; Mccauley, 1984).

2.2. *Microcystin detection classification using snapshot survey*

To evaluate if the coarse temporal resolution of the Ambient Lake Monitoring program influenced the classification of a lake as having detectable microcystin or not in a year, we used a weekly microcystin beach monitoring data set from a subset of lakes ($n = 31$) in this study

(Villanueva et al., 2023) to evaluate classification accuracy. For the beach monitoring program, samples are taken weekly between Memorial Day (end of May) and Labor Day (beginning of September) at public beaches and analyzed for total microcystins concentration in the same manner as the lake monitoring samples. We classified beaches as having detectable microcystin in a given year if one weekly sample was above the detection limit, or non-detectable if no samples exceeded the reporting limit of 0.6 μg L^{-1}. We then compared if the detection classification was the same at the beach as the deep site for the same lake in the same year.

2.3. Classification model development

We used conditional inference forests (Strobl et al., 2007), a machine learning technique, to build a predictive model of microcystin detection from the Ambient Lake Monitoring data set (snapshot survey) as a function of lake physical and chemical attributes, watershed characteristics, zooplankton abundance, and weather. Conditional inference forests (Strobl et al., 2007) are an extension of random forests (Breiman, 2001) designed to overcome particular weaknesses of the traditional random forest algorithm. Like random forests, conditional inference forests build ensembles of tree-based classifiers using random subsets of predictor variables, and share their strengths in dealing with large numbers of predictor variables that may be correlated and have complex interactive effects on the response (Strobl et al., 2009, 2007). Traditional random forests are an ensemble of classification trees (alternatively, regression trees for continuous response variables), which are prone to over-fitting and are biased toward selecting variables for which many splits are possible (i.e., continuous variables or multi-level factors) (Strobl et al., 2007). Conditional inference forests replace classification trees with conditional inference trees, a method overcoming both these weaknesses, resulting in greater parsimony and unbiased variable selection (Hothorn et al., 2006; Strobl et al., 2007). More specifically, classification trees tend to overfit in the sense of making "deep" trees having extra splits in the data that contribute little additional explanatory power, but conditional inference trees implement an explicit test of whether further splits contribute significant explanatory power, thus limiting overfitting. Conditional inference forests were implemented in R version 4.2.1 using the 'party' package (Hothorn, 2005; Strobl et al., 2008, 2007).

We predicted microcystin detection as a function of 37 predictor variables (Appendix S1: Table S1, Fig. 2). Predictors included lake physical attributes (e.g., surface area, stratification), lake chemical measurements (e.g., nutrient concentrations), watershed characteristics (e.g., land cover composition, ecoregion), and zooplankton abundance. Variables from field sampling were annualized by averaging across the three measurements taken each summer. Land cover data for each watershed was provided by the Iowa Department of Natural Resources and ecoregions were defined by the level IV ecoregions from the US Environmental Protection Agency. Two weather variables were computed using PRISM monthly gridded (4 km × 4 km) climate data: total precipitation from October of the prior year through May of the focal year (e.g., October 2016 to May 2017 for microcystin detections in 2017), and mean springtime (March through May) temperature in the current year (PRISM Climate Group, Oregon State University, 2014). The concentration of phycocyanin, a pigment found in cyanobacteria, was excluded from the set of candidate variables because of its obvious association with detection of microcystin. Alternate model results with phycocyanin included in the predictor set are presented in Appendix S1: Figs. S4, S5. The forest consisted of 30,000 trees, which ensured stability of model performance and of the rank-order of absolute variable importance of predictors.

2.4. Model evaluation

We first built a model trained on data from 2017 (n = 128 lakes), evaluated its overall performance, quantified which variables contributed the most explanatory power, and assessed the effects of the most important variables on the probability of microcystin detection. We evaluated model performance using the area under the curve (AUC) of a receiver-operator characteristic (ROC) curve, a widely used metric that, in its application here, balances the model's ability to correctly classify lakes with microcystin detection with its ability to identify lakes where microcystin was not detected. AUC ranges from 0 to 1 with 0.5 indicating a model no better than random assignment and 1 indicating a perfect model. We considered a prediction effective if AUC ≥ 0.7. We computed absolute variable importances and selected the 6 most important predictors for further investigation. We evaluated these variables' partial dependences to assess their effect on microcystin detection using the 'pdp' R library (Greenwell, 2017). The model trained on 2017 data was applied forward in time to predict microcystin detection in 2018–2021 and model performance was assessed using AUC. The number of lakes included for years 2018–2021 were 127, 129, 67 (in 2020 a smaller number of lakes were sampled due to the COVID-19 pandemic disruptions), and 128.

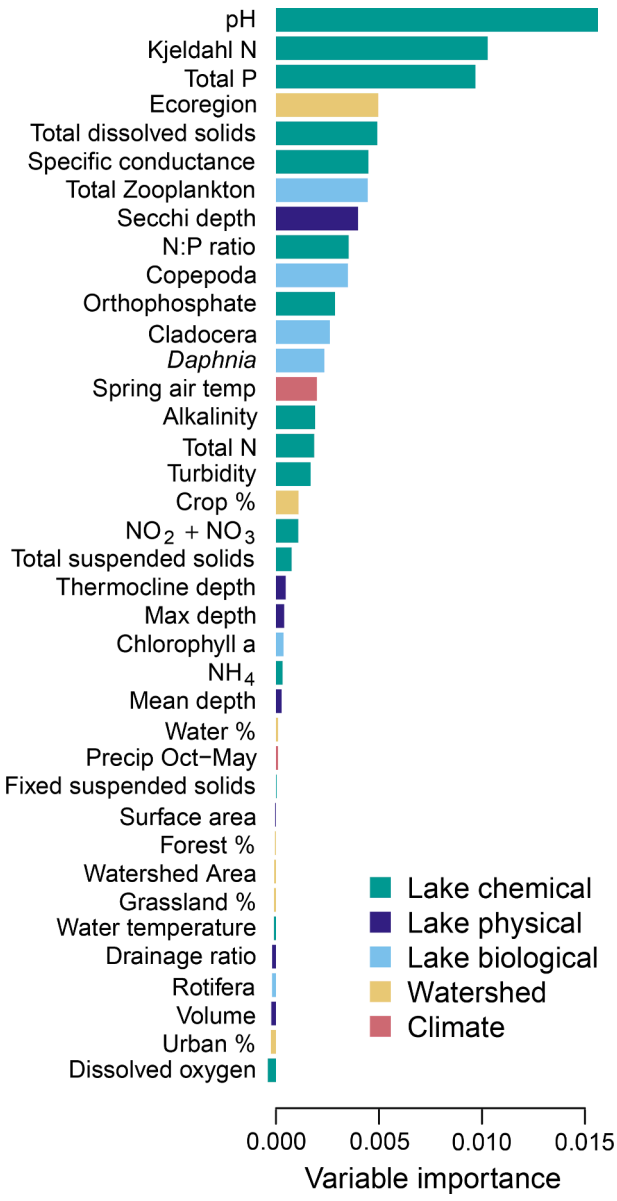


Fig. 2. Variable importance values for predictors for the model trained on 2017 data.

While model performance was assessed primarily using AUC, we also produced confusion matrices for these predictions to further evaluate model performance (Table S2). An optimal threshold on the predicted probability of microcystin detection for assigning predictions to detect and non-detect classes was determined using Youden's Index (Youden, 1950). Youden's Index (*J*) is the sum of sensitivity and specificity minus one, and by balancing sensitivity and specificity it is consistent with our use of AUC as an evaluation metric. We also calculated total accuracy based on the confusion matrices by summing the number of lakes that were true positives (microcystin detection predicted, microcystin detection observed) and true negatives (no detection predicted, no detection observed) and dividing by the total number of lakes sampled that year.

We next tested whether accumulating training data through time (e.g., the model used to predict 2019 microcystin detections was trained on data from 2017 to 2018) improved model performance. Accumulating training data as it's collected is a tenet of iterative near-term forecasting (Dietze, 2017) and previously has been shown to improve model performance in applications to different questions than we focus on here (Carey et al., 2022). As above, model performance was evaluated using AUC.

Finally, we tested the skill of our trained models at predicting microcystin detections later in the season using only field data from the first sampling event of the summer (mid-May through end of June) and variables not derived from field sampling (which were temporally invariant except for the weather variables). Cyanotoxin-producing blooms are most prevalent in late summer, so the ability of our model to predict late-season microcystin detection from early-season observations corresponds to the potential for the combination of a trained model and early-season limnological observations to serve as an early warning system. For this experiment, trained models accumulated observations from subsequent years as they became available (e.g., the model used to predict 2019 microcystin detections was trained on data from 2017 to 2018).

3. Results

There was substantial variability in the number of lakes with microcystin detected and concentration among years. Each year, microcystin was detected above the project reporting limit of 0.6 μg L^{-1} in 37.9–59.8 % of lakes (Fig. 1A). Microcystin concentrations varied among lakes and years, with only 2–3 lakes with mean summer concentrations above the recreational limit of 8 μg L^{-1} in 2020 and 2021, whereas 24 lakes were above this threshold in 2019. There was substantial variability in precipitation among years. Cumulative water year precipitation was 2.4 times higher in 2018 compared to 2021 (Appendix S1: Fig. S1). The classification of lakes as having detectable versus non-detectable microcystin was dynamic. Every year, 19 to 36 lakes changed classification from the previous year. Finally, 81 % of lake-years had the same classification of detectable or non-detectable microcystin when comparing between the beach monitoring program and the lake monitoring program (Appendix S1: Fig. S2). Additionally, 11 % of lake-years did not agree in detection classification between the two monitoring programs, because of a transient period of microcystin detection at the beach (e.g., only 1 week), with no additional detections that summer.

The AUC for the model trained on 2017 data when classifying the same 2017 microcystin detections was 0.972 (95 % CI: 0.950–0.993). Based on the confusion matrix, total accuracy of the 2017 model was 88 % (Table S2). The six most important predictors of microcystin detection in this model were pH, total Kjeldahl N, total P, total dissolved solids, ecoregion, and total zooplankton (Fig. 2). In general, in-lake chemical variables were better predictors than other classes of variables, but the relatively high importance of ecoregion and total zooplankton demonstrate a role for eco-geographic contexts and food web structure in harmful algal bloom dynamics. All continuous variables in the top predictors had threshold-like relationships with the probability of microcystin detection. Threshold values from the partial dependence plots were assessed visually and reported here. The probability of microcystin detection increased with increasing pH (threshold between 8.4–8.6), total Kjeldahl N (threshold 0.6–1.7 mg L^{-1}), total P (threshold 30–70 μg L^{-1}), and total zooplankton (threshold 5–155 mg L^{-1}) (Fig. 3). Total dissolved solids had a negative relationship with the probability of microcystin detection at a threshold from 253 to 347 mg L^{-1} (Fig. 3d). Microcystin detections were most prevalent in the Des Moines Lobe (ecoregion 47b), steeply rolling loess prairie (ecoregion 47e), and eastern IA and MN drift plains (ecoregion 47c), and lowest in the MO alluvial plain (ecoregion 47d) and western loess hills ecoregions (ecoregion 47 m) (Fig. 3e).

The next six most important variables (Fig. 2) for predicting microcystin detection also align with in-lake variables (molar N:P ratio, orthophosphate, Secchi depth, specific conductance) and plankton (copepod biomass, Cladocera biomass). The probability of detection was substantially higher at a molar N:P ratio <30 and decreased sharply from 30 to 80 (Appendix S1: Fig. S3). Orthophosphate had a similar relationship with the probability of microcystin detection as total P, but the threshold at which probabilities plateaued was lower at 30 μg L^{-1}. Specific conductance had a similar relationship with probability of detection as total dissolved solids (Appendix S1: Fig. S3; threshold 389–535 μS cm^{-1}). Higher Secchi depths (>2 m) were associated with lower probability of detecting microcystin (Appendix S1: Fig. S2). Finally, both copepod and Cladocera biomass had similar relationships as total zooplankton with similar threshold biomasses (Appendix S1: Fig. S3). Overall, weather variables were not important for predicting microcystin detection in the 2017 model.

In general, the model trained on 2017 data made skilled predictions when applied to 2018–2021 conditions (Fig. 4a, Table S2). Importantly, the model was at least moderately successful at classifying lakes that changed class from year to year (53 % to 67 % correct), indicating that the model reflected how changing in-lake conditions affect the probability of microcystin detection. Alternatively, the model could have performed well because the same lakes consistently had microcystin-producing blooms and therefore the model succeeded at distinguishing data signatures of groups of lakes that are unrelated to mechanisms of cyanotoxin production, but this does not explain the performance of our model. When applied to future years, model performance did decline and was somewhat variable across years (Fig. 4a). From 2017 to 2019, AUC declined to 0.841 (95 % CI: 0.773–0.910) and remained near this level and above the effective prediction threshold of AUC > 0.7 through 2021. When we used a model built with a training set that accumulated data over time, model performance for predicting microcystin detection in a lake was modestly improved, primarily for 2020 (Fig. 4a). For the model trained on 2017 data, AUC in 2020 was 0.811 (95 % CI: 0.710–0.912) whereas the model built with accumulated training data increased in AUC for 2020 to 0.851 (95 % CI: 0.763–0.939). While the AUC of model predictions based on early observations in the 2017 model only was lower than those based on all observations in all years, using early-season conditions was effective (AUC ≥ 0.7) in predicting late-season microcystin detections in some but not all years (Fig. 4b). Furthermore, in 2018 and 2021, the quality of predictions based on early-season conditions from that year in the 2017 model rivaled the quality of predictions from the full dataset (95 % CIs overlap point estimates).

We also built a version of the 2017 model with phycocyanin, a pigment found in cyanobacteria, as a potential predictor. In this version of the model, phycocyanin was the most important predictor of microcystin detection in a lake (Appendix S1: Fig. S4), followed by a similar set of predictor variables as the non-phycocyanin model. While phycocyanin was an important variable, its inclusion did not improve model skill. The AUC of the models trained on 2017 data with and without phycocyanin were nearly identical (Fig. S5). Similarly, when the 2017 model with phycocyanin was applied to make predictions in 2018–2021, the skill was unchanged from the non-phycocyanin model.

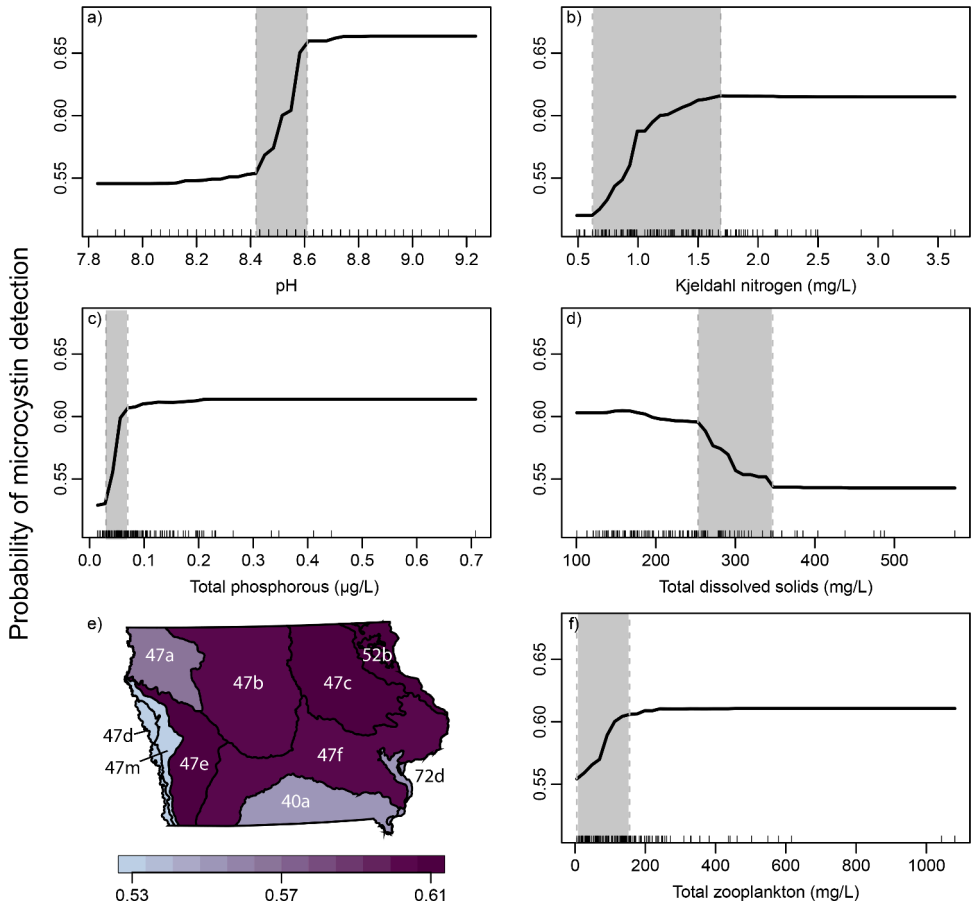


Fig. 3. Partial dependence plots for predictors with the highest variable importance scores in the model trained on 2017 data. A rug plot of the continuous variable (data from 2017) is on the horizontal axis. The ecoregion codes in panel e are described in Figure 1; color gradient in this panel is the probability of microcystin detection (ranging from 0.53 to 0.61, same as the vertical axis in other panels of this figure). Gray polygons indicate the range of values in a variable over which the probability of microcystin detection changed rapidly.

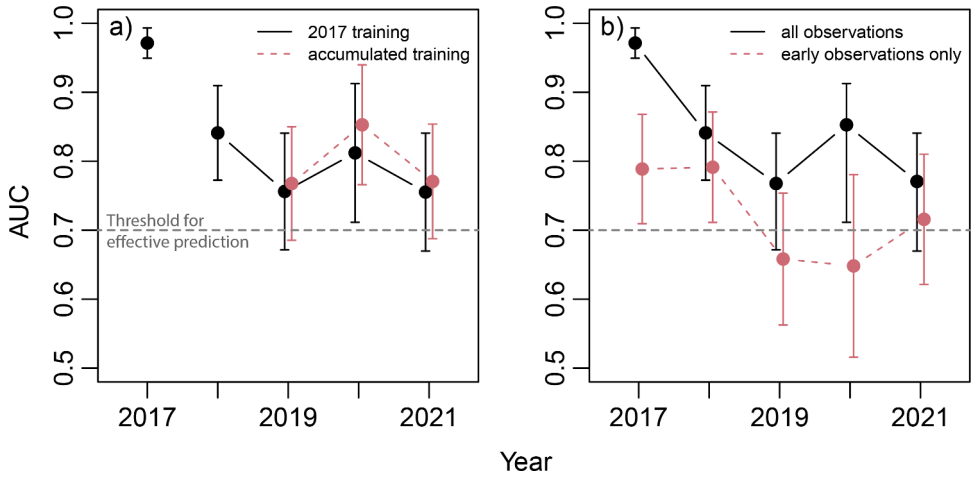


Fig. 4. Model performance (AUC) at classifying microcystin detections across years a) comparing a model trained on 2017 data only to one where the training set accumulates data from all prior years to predict microcystin detections in the next; and b) comparing model performance with predictions based on all observations in a year to predictions using only data from the first sampling event of the season. We considered a prediction effective if AUC $\geq$ 0.7.

4. Discussion

A conditional inference tree model built with snapshot data collected in one year performed well at identifying which lakes had detections of microcystin in subsequent summers. Unlike other snapshot surveys that take advantage of large gradients in nutrient enrichment to statistically identify cyanotoxin-producing conditions, the studied lakes in Iowa are primarily classified as eutrophic and hypereutrophic (i.e., a narrower gradient of enrichment). Regardless, the model was able to differentiate between lakes with and without microcystin present through a combination of productivity, trophic, and eco-region variables. In addition to shedding light on drivers of cyanotoxin production in nutrient-rich

lakes, this exercise demonstrated the utility of developing correlative models based on a single year of data to inform public health decision making, and provided evidence that model performance may not substantially decline under varying climate conditions.

4.1. Drivers of cyanotoxin presence in lakes

The most important variable in the microcystin detection model was pH. When phytoplankton production is high (such as during a bloom), aqueous CO_2 can be depleted and pH increases (Balmer and Downing, 2011; MacKeigan et al., 2023). Cyanobacteria have the ability to overcome this brief CO_2 scarcity through a bicarbonate concentrating mechanism, which helps to maintain productivity (Morales-Williams et al., 2017). There was a substantial increase in microcystin detection between pH 8.4 and 8.6 which coincides with the pH range where CO_2 is a vanishingly small proportion of the inorganic carbon pool (<1 %). This threshold is consistent with the relationship between pH and cyanobacteria biomass observed across a latitudinal gradient in the Americas (Bonilla et al., 2023) and the relationship between pH and hazardous versus safe microcystin concentrations in a subset of Iowa lakes (Villanueva et al., 2023). From this relationship, it is evident that the biogeochemical conditions favoring cyanobacteria dominance and the current degree of productivity during a bloom are important predictors of cyanotoxin production.

Total nutrient concentrations were the next most important predictors, with total P and total Kjeldahl N having positive, saturating relationships with microcystin detection. The importance of these predictors and direction of the relationship is consistent with numerous other cyanobacteria prediction models (Rousso et al., 2020). Iowa is a nutrient-rich landscape with more than 75 % of the lakes having total P concentrations well above the threshold for 'eutrophic' classification. Between 30–60 μg L^{-1} of total P, microcystin detection increased substantially. This range in concentration aligns with the delineation between mesotrophic and eutrophic conditions on the low end (total $P =$ 24 μg L^{-1}; Carlson 1970) to the mid-point between eutrophic and hypereutrophic conditions (total $P = 96$ μg L^{-1}) on the high end. In this nutrient-rich landscape, discrete trophic state classification was not as useful as direct measurements of total P concentrations for predicting cyanotoxin presence (Kraemer, 2020). Increases in total Kjeldahl N from 0 to 1.7 mg L^{-1} were associated with higher detection probability of microcystin. Interestingly, total N, which is effectively a measure of the same pool with the addition of nitrate (Stanley et al., 2019) was not an important variable in the model. In general, inorganic nutrients were not important, which indicates that the measured total nutrient concentrations are likely an index of planktonic biomass, and available inorganic nutrients in concentrations near the model thresholds identified were rapidly taken up to support planktonic production. The molar ratio of total N to total P was a less important predictor, but the threshold over which the probability of detection declined was consistent with the range of molar ratios associated with high microcystin concentrations in other regions (Orihel et al. 2012).

Eco-geographic context of the lake was an important variable in the model indicated by the rank importance of level IV ecoregion and total dissolved solids for the probability of microcystin detection. The ecoregions with the highest probability of microcystin detection coincided with areas of the most extensive cropland in the state, providing a substantial non-point source of nutrients that support cyanobacteria production. These ecoregions tended to have lower concentrations of total dissolved solids. The pattern observed here is consistent with other continental-scale analyses that showed agriculturally-dominated ecoregions were dominated by microcystin detection in lakes (Beaver et al., 2014). Conversely, any ecoregions in Iowa that include substantial pasture and other land cover types besides croplands were less likely to have microcystin detections for lakes in that region. Even in this highly modified landscape, small differences in land use influence the probability of microcystin detection in a waterbody.

Finally, higher zooplankton biomass was associated with a greater probability of detecting microcystin in Iowa lakes, a pattern which has been observed in other regions as well (Ger et al., 2016; MacKeigan et al., 2023). The biomass of N-rich zooplankton taxa such as cyclopoid copepods increase with eutrophication in these lakes, possibly enhancing P availability through excretion (Moody and Wilkinson, 2019) and thereby supporting cyanobacteria production. Copepod biomass was a significant predictor of microcystin presence or absence in a large-scale snapshot survey across Canada (MacKeigan et al., 2023). Selective grazing by zooplankton on non-toxic phytoplankton taxa could support the dominance of microcystin-producing cyanobacteria (Ger et al., 2011; Wang et al., 2010). Both copepods and *Daphnia* have demonstrated selective grazing behavior in the presence of toxic cyanobacteria (Ger et al., 2011; Tillmanns et al., 2011). Alternatively (but perhaps not mutually exclusive), predation release on zooplankton seasonally coinciding with toxic bloom formation may explain the positive relationship between zooplankton and microcystin detection probability. As such, the positive association between zooplankton biomass and probability of microcystin detection may be a direct effect of zooplankton grazing or indirect association between the two variables.

4.2. Model performance and application

Overall and across modeling experiments, we were able to explain and predict microcystin detections in Iowa lakes. Several prior studies have statistically modeled microcystin detections and/or concentrations as a function of lake, watershed and weather variables using snapshot survey data (e.g., Beaver et al., 2014; MacKeigan et al., 2023; Taranu et al., 2017), with the goal of identifying the conditions associated with cyanotoxin detection to inform management. Here, we had the opportunity to evaluate the skill of such predictions from year to year as climate and in-lake conditions changed, a critical component of the continuous improvement process for cyanobacteria prediction models (2020). Our results show that a model of microcystin detection among lakes trained from one year of survey data can be applied forward to predict microcystin detections at a season-wide temporal scale with reasonably high accuracy. In some, but not all, instances model skill can be improved by accumulating data over time as in iterative near-term forecasting (Carey et al., 2022), but in our study this effect was quite modest even though meteorological conditions changed markedly across years (Appendix S1: Fig. S1).

In general, weather variables were not informative predictors of microcystin detection in our study within a single year (Fig. 2) or when including springtime conditions (Fig. 4b). In fact, the model predictions built on springtime conditions alone performed worse overall (although still moderately skilled), compared to all other modeling scenarios. Water temperature is an important predictor in most within-season cyanobacteria bloom forecasting models for a single ecosystem (Rousso et al., 2020). At broader continental scales there is evidence that temperature may drive total cyanobacteria abundance but reduces toxicity (Ho and Michalak, 2020), whereas temperature was not an important predictor of cyanobacteria biomass across larger latitudinal gradients (Bonilla et al., 2023). Similarly, the relationship between precipitation and cyanobacteria abundance at broader spatial scales is uncertain (Ho and Michalak, 2020), whereas extreme precipitation events within a year have been tied to extreme blooms in a lake within the same year (Carpenter et al., 2022b). While meteorological variables may play an important role in microcystin congener dominance (Taranu et al., 2019) and bloom initiation, in our study overall microcystin detection was not influenced by meteorological variability among years (despite variability in precipitation), providing evidence that the statistical models from snapshot surveys can be reliable in future years.

Cyanobacterial blooms are both spatially and temporally dynamic (Buelo et al., 2022; Ortiz and Wilkinson, 2021), which may influence our ability to detect microcystin with the monitoring program design of

sampling at one location (i.e., over the deepest part of the lake), three times per summer. Using a weekly monitoring data set for microcystin at beaches in a subset of the lakes in the study, we found strong agreement (81 % of lake years) between detection of microcystin at the deep site at some point during the summer and detection at the beach. The agreement among these two data sets supports the use of the less frequent monitoring program data in this study to evaluate the performance of a model built on one year of data, applied forward to subsequent years. In other words, it is less likely that model performance, evaluated as skill in categorizing a lake as having detectable microcystin or not each year, was influenced by misclassification in the observational data.

5. Conclusions

Despite the threat to public health, effective cyanotoxin monitoring of recreational waters can be inadequate because it is both resource- and time- intensive to accomplish. Data-driven decision-making tools that help direct limited resources are valuable for resource managers, but only when there is confidence in the tool. Using five years of microcystin and water quality monitoring data from lakes in a nutrient-rich, highly modified landscape, we were able to evaluate the utility and accuracy of correlative models built on a single year's worth of data applied to future years. While the performance of the 2017 model declined when applied to future years, model skill remained relatively high despite lakes shuffling microcystin detection categories each year and a narrow range of trophic state classifications. Our results support the application of correlative analyses built on single-season sampling data to decision-making for resource allocation, but we call for similar investigations in other regions to build further evidence for this approach.

CRediT authorship contribution statement

Grace M. Wilkinson: Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Jonathan A. Walter:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Ellen A. Albright:** Writing – review & editing, Investigation, Conceptualization. **Rachel F. King:** Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Eric K. Moody:** Writing – review & editing, Investigation, Conceptualization. **David A. Ortiz:** Writing – review & editing, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Collated data sets and analysis scripts used for this study are available at https://doi.org/10.5281/zenodo.12005095.

Acknowledgements

We thank the many technicians that contributed to data collection and sample analysis as a part of the Ambient Lake Monitoring Program at Iowa State University.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.hal.2024.102679.

References

Aboal, M., Puig, M.Á., Asencio, A.D., 2005. Production of microcystins in calcareous Mediterranean streams: the Alharabe River, Segura River basin in south-east Spain. J. Appl. Phycol. 17, 231–243. https://doi.org/10.1007/s10811-005-2999-z.

Arbuckle, K.E., Downing, J.A., 2001. The influence of watershed land use on lake N: p in a predominantly agricultural landscape. Limnol. Oceanogr. 46, 970–975. https://doi.org/10.4319/lo.2001.46.4.0970.

Balmer, M., Downing, J., 2011. Carbon dioxide concentrations in eutrophic lakes: undersaturation implies atmospheric uptake. Inland Waters 1, 125–132. https://doi.org/10.5268/IW-1.2.366.

Beal, M.R.W., Wilkinson, G.M., Block, P.J., 2023. Large scale seasonal forecasting of peak season algae metrics in the Midwest and Northeast U.S. Water Res. 229, 119402 https://doi.org/10.1016/j.watres.2022.119402.

Beaver, J.R., Manis, E.E., Loftin, K.A., Graham, J.L., Pollard, A.I., Mitchell, R.M., 2014. Land use patterns, ecoregion, and microcystin relationships in U.S. lakes and reservoirs: a preliminary evaluation. Harmful Algae 36, 57–62. https://doi.org/10.1016/j.hal.2014.03.005.

Beversdorf, L.J., Miller, T.R., McMahon, K.D., 2015. Long-term monitoring reveals carbonâ€“nitrogen metabolism key to microcystin production in eutrophic lakes. Front. Microbiol. 6 https://doi.org/10.3389/fmicb.2015.00456.

Bonilla, S., Aguilera, A., Aubriot, L., Huszar, V., Almanza, V., Haakonsson, S., Izaguirre, I., O'Farrell, I., Salazar, A., Becker, V., Cremella, B., Ferragut, C., Hernandez, E., Palacio, H., Rodrigues, L.C., Sampaio Da Silva, L.H., Santana, L.M., Santos, J., Somma, A., Ortega, L., Antoniades, D., 2023. Nutrients and not temperature are the key drivers for cyanobacterial biomass in the Americas. Harmful Algae 121, 102367. https://doi.org/10.1016/j.hal.2022.102367.

Breiman, L., 2001. Random forest. Mach. Learn. 25, 5–32. https://doi.org/10.1023/A:1010950718922.

Buelo, C.D., Pace, M.L., Carpenter, S.R., Stanley, E.H., Ortiz, D.A., Ha, D.T., 2022. Evaluating the performance of temporal and spatial early warning statistics of algal blooms. Ecol. Appl. 32 https://doi.org/10.1002/eap.2616.

Carey, C.C., Woelmer, W.M., Lofton, M.E., Figueiredo, R.J., Bookout, B.J., Corrigan, R.S., Daneshmand, V., Hounshell, A.G., Howard, D.W., Lewis, A.S.L., McClure, R.P., Wander, H.L., Ward, N.K., Thomas, R.Q., 2022. Advancing lake and reservoir water quality management with near-term, iterative ecological forecasting. Inland Waters 12, 107–120. https://doi.org/10.1080/20442041.2020.1816421.

Carmichael, W.W., Boyer, G.L., 2016. Health impacts from cyanobacteria harmful algae blooms: implications for the North American Great Lakes. Harmful Algae 54, 194–212. https://doi.org/10.1016/j.hal.2016.02.002.

Carpenter, S.R., Arani, B.M.S., Van Nes, E.H., Scheffer, M., Pace, M.L., 2022a. Resilience of phytoplankton dynamics to trophic cascades and nutrient enrichment. Limnol. Oceanogr. 67 https://doi.org/10.1002/lno.11913.

Carpenter, S.R., Gahler, M.R., Kucharik, C.J., Stanley, E.H., 2022b. Long-range dependence and extreme values of precipitation, phosphorus load, and Cyanobacteria. Proc. Natl. Acad. Sci. U. S. A. 119, e2214343119 https://doi.org/10.1073/pnas.2214343119.

Clark, J.S., Carpenter, S.R., Barber, M., Collins, S., Dobson, A., Foley, J.A., Lodge, D.M., Pascual, M., Pielke, R., Pizer, W., Pringle, C., Reid, W.V., Rose, K.A., Sala, O., Schlesinger, W.H., Wall, D.H., Wear, D., 2001. Ecological forecasts: an emerging imperative. Science 293, 657–660. https://doi.org/10.1126/science.293.5530.657 (1979).

de Figueiredo, D.R., Azeiteiro, U.M., Esteves, S.M., Gonçalves, F.J.M., Pereira, M.J., 2004. Microcystin-producing blooms—A serious global public health issue. Ecotoxicol. Environ. Saf. 59, 151–163. https://doi.org/10.1016/j.ecoenv.2004.04.006.

Dietze, M.C., 2017. Prediction in ecology: a first-principles framework. Ecol. Appl. 27, 2048–2060. https://doi.org/10.1002/eap.1589.

Dietze, M.C., Fox, A., Beck-Johnson, L.M., Betancourt, J.L., Hooten, M.B., Jarnevich, C.S., Keitt, T.H., Kenney, M.A., Laney, C.M., Larsen, L.G., Loescher, H.W., Lunch, C.K., Pijanowski, B.C., Randerson, J.T., Read, E.K., Tredennick, A.T., Vargas, R., Weathers, K.C., White, E.P., 2018. Iterative near-term ecological forecasting: needs, opportunities, and challenges. Proc. Natl. Acad. Sci. U. S. A. 115, 1424–1432. https://doi.org/10.1073/pnas.1710231115.

Dumont, H.J., Van de Velde, I., Dumont, S., 1975. The dry weight estimate of biomass in a selection of Cladocera, Copepoda and Rotifera from the plankton, periphyton and benthos of continental waters. Oecologia 19, 75–97. https://doi.org/10.1007/BF00377592.

Filstrup, C.T., Downing, J.A., 2017. Relationship of chlorophyll to phosphorus and nitrogen in nutrient-rich lakes. Inland Waters 7, 385–400. https://doi.org/10.1080/20442041.2017.1375176.

Filstrup, C.T., Hillebrand, H., Heathcote, A.J., Harpole, W.S., Downing, J.A., 2014. Cyanobacteria dominance influences resource use efficiency and community turnover in phytoplankton and zooplankton communities. Ecol. Lett. 17, 464–474. https://doi.org/10.1111/ele.12246.

Ger, K.A., Panosso, R., Lürling, M., 2011. Consequences of acclimation to *Microcystis* on the selective feeding behavior of the calanoid copepod *Eudiaptomus gracilis*. Limnol. Oceanogr. 56, 2103–2114. https://doi.org/10.4319/lo.2011.56.6.2103.

Ger, K.A., Urrutia-Cordero, P., Frost, P.C., Hansson, L.A., Sarnelle, O., Wilson, A.E., Lürling, M., 2016. The interaction between cyanobacteria and zooplankton in a more eutrophic world. Harmful Algae 54, 128–144. https://doi.org/10.1016/j.hal.2015.12.005.

Gobler, C.J., Burkholder, J.M., Davis, T.W., Harke, M.J., Johengen, T., Stow, C.A., Van De Waal, D.B., 2016. The dual role of nitrogen supply in controlling the growth and toxicity of cyanobacterial blooms. Harmful Algae 54, 87–97. https://doi.org/10.1016/j.hal.2016.01.010.

Gorney, R.M., June, S.G., Stainbrook, K.M., Smith, A.J., 2023. Detections of cyanobacteria harmful algal blooms (cyanoHABs) in New York State, United States (2012–2020). Lake Reserv. Manage 39, 21–36. https://doi.org/10.1080/10402381.2022.2161436.

Greenwell, B.M., 2017. pdp: an R package for constructing partial dependence plots. R J. 9, 421. https://doi.org/10.32614/RJ-2017-016.

Harris, T.D., Wilhelm, F.M., Graham, J.L., Loftin, K.A., 2014. Experimental manipulation of TN:TP ratios suppress cyanobacterial biovolume and microcystin concentration in large-scale *in situ* mesocosms. Lake Reserv. Manag. 30, 72–83. https://doi.org/10.1080/10402381.2013.876131.

Hillebrand, H., Dürselen, C., Kirschtel, D., Pollingher, U., Zohary, T., 1999. Biovolume calculation for pelagic and benthic microalgae. J. Phycol. 35, 403–424. https://doi.org/10.1046/j.1529-8817.1999.3520403.x.

Ho, J.C., Michalak, A.M., 2020. Exploring temperature and precipitation impacts on harmful algal blooms across continental U.S. lakes. Limnol. Oceanogr. 65, 992–1009. https://doi.org/10.1002/lno.11365.

Ho, J.C., Michalak, A.M., Pahlevan, N., 2019. Widespread global increase in intense lake phytoplankton blooms since the 1980s. Nature 574, 667–670. https://doi.org/10.1038/s41586-019-1648-7.

Holmes, R., Norris, R., Smayda, T., Wood, E., 1969. Collection, fixation, identification, and enumeration of phytoplankton standing stock. Recommended Procedures for Measuring the Productivity of Plankton Standing Stock and Related Oceanic Properties. National Academy of Sciences.

Hothorn, T., 2005. Survival ensembles. Biostatistics 7, 355–373. https://doi.org/10.1093/biostatistics/kxj011.

Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning: a conditional inference framework. J. Comput. Graph. Stat. 15, 651–674. https://doi.org/10.1198/106186006X133933.

Kincaid, D.W., Seybold, E.C., Adair, E.C., Bowden, W.B., Perdrial, J.N., Vaughan, M.C.H., Schroth, A.W., 2020. Land use and season influence event-scale nitrate and soluble reactive phosphorus exports and export stoichiometry from headwater catchments. Water Resour. Res. 56, e2020WR027361 https://doi.org/10.1029/2020WR027361.

Kraemer, B.M., 2020. Rethinking discretization to advance limnology amid the ongoing information explosion. Water Res. 178, 115801 https://doi.org/10.1016/j.watres.2020.115801.

Loecke, T.D., Burgin, A.J., Riveros-Iregui, D.A., Ward, A.S., Thomas, S.A., Davis, C.A., Clair, M.A.St., 2017. Weather whiplash in agricultural regions drives deterioration of water quality. Biogeochemistry 133, 7–15. https://doi.org/10.1007/s10533-017-0315-z.

Lofton, M.E., Howard, D.W., Thomas, R.Q., Carey, C.C., 2023. Progress and opportunities in advancing near-term forecasting of freshwater quality. Glob. Change Biol. 29, 1691–1714. https://doi.org/10.1111/gcb.16590.

MacKeigan, P.W., Zastepa, A., Taranu, Z.E., Westrick, J.A., Liang, A., Pick, F.R., Beisner, B.E., Gregory-Eaves, I., 2023. Microcystin concentrations and congener composition in relation to environmental variables across 440 north-temperate and boreal lakes. Sci. Total Environ. 884, 163811 https://doi.org/10.1016/j.scitotenv.2023.163811.

Mccauley, E., 1984. The estimation of the abundance and biomass of zooplankton in samples.

Moody, E.K., Wilkinson, G.M., 2019. Functional shifts in lake zooplankton communities with hypereutrophication. Freshw. Biol. 64, 608–616. https://doi.org/10.1111/fwb.13246.

Morales-Williams, A.M., Wanamaker Jr., A.D., Downing, J.A., 2017. Cyanobacterial carbon concentrating mechanisms facilitate sustained CO₂ depletion in eutrophic lakes. Biogeosciences 14, 2865–2875. https://doi.org/10.5194/bg-14-2865-2017.

Orihel, D.M., Bird, D.F., Brylinsky, M., Chen, H., Donald, D.B., Huang, D.Y., Giani, A., Kinniburgh, D., Kling, H., Kotak, B.G., Leavitt, P.R., Nielsen, C.C., Reedyk, S., Rooney, R.C., Watson, S.B., Zurawell, R.W., Vinebrooke, R.D., 2012. High microcystin concentrations occur only at low nitrogen-to-phosphorus ratios in nutrient-rich Canadian lakes. Can. J. Fish. Aquat. Sci. 69, 1457–1462. https://doi.org/10.1139/f2012-088.

Ortiz, D.A., Wilkinson, G.M., 2021. Capturing the spatial variability of algal bloom development in a shallow temperate lake. Freshw. Biol. 66, 2064–2075. https://doi.org/10.1111/fwb.13814.

PRISM Climate Group, Oregon State University, 2014. https://prism.oregonstate.edu.

Rastogi, R.P., Sinha, R.P., Incharoensakdi, A., 2014. The cyanotoxin-microcystins: current overview. Rev. Environ. Sci. Biotechnol. 13, 215–249. https://doi.org/10.1007/s11157-014-9334-6.

Reinl, K.L., Harris, T.D., North, R.L., Almela, P., Berger, S.A., Bizic, M., Burnet, S.H., Grossart, H., Ibelings, B.W., Jakobsson, E., Knoll, L.B., Lafrancois, B.M., McElarney, Y., Morales-Williams, A.M., Obertegger, U., Ogashawara, I., Paule-Mercado, M.C., Peierls, B.L., Rusak, J.A., Sarkar, S., Sharma, S., Trout-Haney, J.V., Urrutia-Cordero, P., Venkiteswaran, J.J., Wain, D.J., Warner, K., Weyhenmeyer, G. A., Yokota, K., 2023. Blooms also like it cold. Limnol. Oceanogr. Lett. 8, 546–564. https://doi.org/10.1002/lol2.10316.

Rousso, B.Z., Bertone, E., Stewart, R., Hamilton, D.P., 2020. A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes. Water Res. 182, 115959 https://doi.org/10.1016/j.watres.2020.115959.

Shingai, Q.K., Wilkinson, G.M., 2023. Microcystin as a biogeochemical cycle: pools, fluxes, and fates of the cyanotoxin in inland waters. Limnol. Oceanogr. Lett. 8, 406–418. https://doi.org/10.1002/lol2.10300.

Stanley, E.H., Rojas-Salazar, S., Lottig, N.R., Schliep, E.M., Filstrup, C.T., Collins, S.M., 2019. Comparison of total nitrogen data from direct and Kjeldahl-based approaches in integrated data sets. Limnol. Oceanogr. Methods 17, 639–649. https://doi.org/10.1002/lom3.10338.

Stoddard, J.L., Van Sickle, J., Herlihy, A.T., Brahney, J., Paulsen, S., Peck, D.V., Mitchell, R., Pollard, A.I., 2016. Continental-scale increase in lake and stream phosphorus: are oligotrophic systems disappearing in the United States? Environ. Sci. Technol. 50, 3409–3415. https://doi.org/10.1021/acs.est.5b05950.

Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinform. 9, 307. https://doi.org/10.1186/1471-2105-9-307.

Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinform. 8, 25. https://doi.org/10.1186/1471-2105-8-25.

Strobl, C., Malley, J., Tutz, G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol. Methods 14, 323–348. https://doi.org/10.1037/a0016973.

Taranu, Z.E., Gregory-Eaves, I., Steele, R.J., Beaulieu, M., Legendre, P., 2017. Predicting microcystin concentrations in lakes and reservoirs at a continental scale: a new framework for modelling an important health risk factor: TARANU et al. Glob. Ecol. Biogeogr. 26, 625–637. https://doi.org/10.1111/geb.12569.

Taranu, Z.E., Pick, F.R., Creed, I.F., Zastepa, A., Watson, S.B., 2019. Meteorological and nutrient conditions influence microcystin congeners in freshwaters. Toxins 11, 620. https://doi.org/10.3390/toxins11110620 (Basel).

Tillmanns, A.R., Burton, S.K., Pick, F.R., 2011. Daphnia pre-exposed to toxic microcystis exhibit feeding selectivity. Int. Rev. Hydrobiol. 96, 20–28. https://doi.org/10.1002/iroh.201011298.

Van De Waal, D.B., Smith, V.H., Declerck, S.A.J., Stam, E.C.M., Elser, J.J., 2014. Stoichiometric regulation of phytoplankton toxins. Ecol. Lett. 17, 736–742. https://doi.org/10.1111/ele.12280.

Villanueva, P., Yang, J., Radmer, L., Liang, X., Leung, T., Ikuma, K., Swanner, E.D., Howe, A., Lee, J., 2023. One-week-ahead prediction of cyanobacterial harmful algal blooms in Iowa lakes. Environ. Sci. Technol. 3c07764. https://doi.org/10.1021/acs.est.3c07764 acs.est.

Wagner, N.D., Quach, E., Buscho, S., Ricciardelli, A., Kannan, A., Naung, S.W., Phillip, G., Sheppard, B., Ferguson, L., Allen, A., Sharon, C., Duke, J.R., Taylor, R.B., Austin, B.J., Stovall, J.K., Haggard, B.E., Chambliss, C.K., Brooks, B.W., Scott, J.T., 2021. Nitrogen form, concentration, and micronutrient availability affect microcystin production in cyanobacterial blooms. Harmful Algae 103, 102002. https://doi.org/10.1016/j.hal.2021.102002.

Walter, J.A., Grage, K., Nunez-Mir, G.C., Grayson, K.L., 2023. Forecasting the spread of an invasive forest-defoliating insect. Divers. Distrib. 13799. https://doi.org/10.1111/ddi.13799 ddi.

Wang, X., Qin, B., Gao, G., Paerl, H.W., 2010. Nutrient enrichment and selective predation by zooplankton promote Microcystis (Cyanobacteria) bloom formation. J. Plankton Res. 32, 457–470. https://doi.org/10.1093/plankt/fbp143.

Wheeler, K.I., Dietze, M.C., LeBauer, D., Peters, J.A., Richardson, A.D., Ross, A.A., Thomas, R.Q., Zhu, K., Bhat, U., Munch, S., Buzbee, R.F., Chen, M., Goldstein, B., Guo, J., Hao, D., Jones, C., Kelly-Fair, M., Liu, H., Malmborg, C., Neupane, N., Pal, D., Shirey, V., Song, Y., Steen, M., Vance, E.A., Woelmer, W.M., Wynne, J.H., Zachmann, L., 2024. Predicting spring phenology in deciduous broadleaf forests: NEON phenology forecasting community challenge. Agric. For. Meteorol. 345, 109810 https://doi.org/10.1016/j.agrformet.2023.109810.

Wilkinson, G.M., Walter, J.A., Buelo, C.D., Pace, M.L., 2022. No evidence of widespread algal bloom intensification in hundreds of lakes. Front. Ecol. Environ. 20, 16–21. https://doi.org/10.1002/fee.2421.

Youden, W., 1950. Index for rating diagnostic tests. Cancer 3, 32–35.